The key difference between the two lies in how they update the Q-values: Q-learning uses the maximum Q-value of the next state, while SARSA uses the Q-value of the actual next action based on the current policy.

Q-learning often tends to be more explorative early on because it's always considering the best possible next action (greedy) regardless of the policy it's currently following. This can lead to riskier moves.

SARSA, being on-policy, tends to be more conservative. Especially in scenarios with potential penalties, SARSA's policy might steer clear of areas with high negative rewards even if it's the faster route to the goal, especially when ε is high.

- Q-learning might appear faster in reaching the goal since it aims for the highest rewards.
- SARSA might seem slower, especially if it's taking a more roundabout or conservative path to avoid potential penalties.

SARSA might take more episodes initially to reach the goal consistently.
Q-learning might reach the goal in fewer episodes, but it might also occasionally fail if it takes riskier paths.